ELSEVIER

Journal of Alloys and Compounds 279 (1998) 14–21

Journal of ALLOYS AND COMPOUNDS

Rough sets applied to the discovery of materials knowledge

A.G. Jackson[a,*], Z. Pawlak[b], S.R. LeClair[a]

[a]*Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright Patterson Air Force Base, Ohio, USA*
[b]*Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, ul. Baltycka 5, 44 100 Gliwice, Poland*

Abstract

The functional mapping of material structure to properties, processing, and use is the principal driver for all scientific and engineering endeavors. Given the high cost of experimentation and the computational intractability of ab initio materials research, more efficient and accurate predictions of yet-to-be-made materials is an equally prominent endeavor, if not a preeminent materials research frontier. Because of the vast amounts of information to be considered in the pursuit of either, the automation of search-based methods for augmenting more analytic approaches is receiving increasing attention. Given the computational challenges to automation and to retrieving data from complex databases, search-based methods offer an expeditious approach to providing a researcher both insight and perspective. Rough sets is discussed relative to these objectives, as is current research to address its limitations and difficulties in application. Several materials related examples are offered to illustrate the application of the method.

Keywords: Material design; Rough sets; Pyramidal networks

1. Introduction

Materials, and therein materials design, has evolved over two millennia from the macromonolithic and composite materials toward the ‘atomic-scale’ control of lattices, surfaces and interfaces sometimes referred to as ‘crystal engineering’, [1] the intricacies and difficulties of which are most notably manifested in the material processing associated with their manufacture.

Currently, crystal engineering involves a sundry of empirically-driven approaches. Although a more math-theoretic approach is desirable, computational intractability precludes their practical application. Even with the advent of adequate computational speed, the marketplace will invariably reward those methods which yield profitable results and are refined autonomously via experimental and/or empirical data. Such methods will need to be used in conjunction with and/or augment, or when necessary. supplant the below computational approaches for a lack of fundamental theory:

1. large scale 1st principle calculations,
   e.g., density functional theory as applied to multi-element property calculations,
2. global optimization of ‘*n*’ dimensional materials design,
3. large scale molecular dynamics applied to property prediction,
   e.g., defects as they relate to microstructure and related properties,
4. atomic-scale (10^{-13} seconds) structure evolution during material processing,
   e.g., epitaxial growth of thin film semiconductor materials.

The development of a theory to support atomic-scale materials design may take decades, and when available it may still be impractical at the desired scale. An example, relative to the above ‘atomic-level’ structure evolution, illustrates both a research goal in materials design and an opportunity for new search-based methods in the process modeling and growth control of thin-film epitaxy.

The ultimate goal of semiconductor thin-film epitaxy is the ability to produce structures where every atom is exactly positioned in its designed location. This would allow for the greatest degree of semiconductor device miniaturization possible (e.g., single electron transistors, CRAY supercomputers on a single chip). A more realistic near term goal is the ability to produce atomically abrupt interfaces between different materials (e.g., GaAs and AlGaAs). This would enable the production of optimal quantum effect devices, leading to revolutionary advances for a host of electronic and optical applications.

*Corresponding author.

0925-8388/98/$19.00
PII: S0925-8388(98)00607-0

Achieving the level of control required for atomically abrupt interfaces is a considerable challenge. The state-of-the-art process for producing these heterostructures is molecular beam epitaxy (MBE). MBE currently suffers from a lack of reproducibility and from limited real-time in situ sensing. Typically, prolonged trial-and-error experiments are required to develop such new material structures and to both develop and/or validate the theory of 'atomic-level' processing physics.

Theoretical models aid in understanding growth processes at the atomic-scale, but due to the disparate time scales associated with important growth parameters, at least two models are required. One, the atomic-scale ab initio calculations, can provide the surface structure and the physical mechanisms (with corresponding activation energy barriers) required by the other model. But the other more macro-scale model, effusion cell growth, is based upon process geometry and associated constraints involving variations in temperatures, and therein flux pressures which occur at 10^{-3} seconds resolution. Consequently, ab initio calculations at the 10^{-13} seconds cannot be used for in situ models of the actual growth (one monolayer per second), because the disparate time scales would require 10^{13} computationally intense time steps per simulated layer grown. For any reasonable surface area each time step, ab initio calculation, may typically require 1 s to compute. Hence, one monolayer of growth would take over 300 000 years to simulate. Although more accurate, computationally viable ab initio algorithms for the dynamics and energetics of semiconductor surfaces need to be developed, the key modeling research question is, how do we effectively link the two time scales [2]?

Perusal of the literature [3] suggests linking of the time scales may be both achievable and more tractable via cellular automata concepts, i.e., in lieu of massively parallel hardware, a massively parallel software implementation of a finite state machine. Our research, on the use of cellular automata concepts applied to the modeling of thin film deposition, represents an approach to the problem of atomistic modeling of deposition processes based on the application of discretized linear models or rules. The use of rules, such as depicted in Fig. 1, to model a deposition process (i.e., the film growth behavior) is comparable to ballistic and random deposition models involving complex differential equations. The simplicity afforded using rules in conjunction with a parallelized software implementation of a finite state machine enables a more tractable modeling of complex film growth/behavior, particularly in three dimensions.

Because these process rules are based upon empirical observations, a second challenge and the principal focus of this paper, is the discovery of these rules for new materials/processes and their subsequent refinement to more accurately represent growth processes. If growth processes such as MBE were augmented with real-time data acquisition involving in situ sensing of substrate temperature, film composition, flux density, and thickness, then the requisite data would be available to link the two time scales. Although, initially, the sensed data would be limited to a representative set of points on the wafer surface, eventually data acquisition and processing speed would enable a detailed two-dimensional (2D) view of each parameter across the wafer surface.

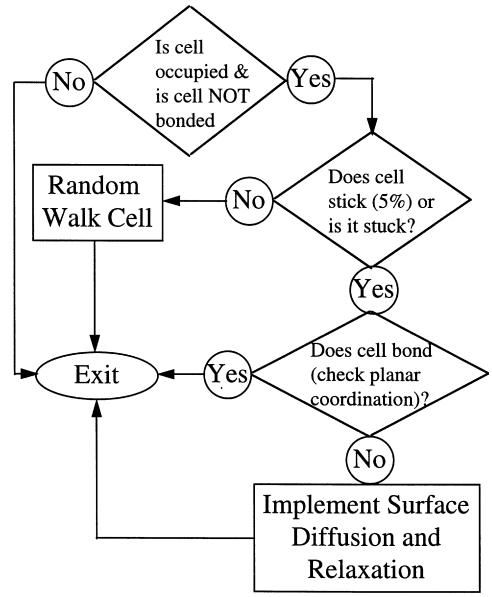

Fig. 1. Cellular automata thin film growth rule.

2. Structure–property–process relations

To bridge the gap in computational tractability, researchers are using visualization tools for understanding increasingly complex data. In fact, data acquisition and analysis (software) applications have even exceeded the use of computer-aided design (CAD), mechanical modeling, image analysis, and mathematical function applications [4]. As a consequence, it is expected that more efficient and visually stimulating search-based methods will be sought to view and compare data, e.g., multiple growth runs (10–20 minute growths at 1 s increments) of a specific recipe with data resolution limited to discernible step-changes in sensed parameters. Successive increases in resolution will be automatically evaluated, looking for patterns in the data which identify regularities and/or irregularities. Once these characteristic patterns are identified, still greater resolution (up to 10^{-3} seconds today and projected to improve to 10^{-13} seconds as necessary) will be used for correlating film and growth process parameters. Eventually these patterns will be compared in near real-time with ab initio models and, in time, begin to close the gap between new materials model development and validation.

In order to compare these characteristic dynamic processing patterns at 10^{-13} or even 10^{-1} seconds, search-based methods of large static data sets will need to be extended and automated. This work has already begun as applied to existing and moderately large data sets of material structure-property-process data [5–7]. Such 'mining' of large static data sets is but the first step toward the frontier of on-line property-process discovery. Not only will the temporal order and volume of process data present new challenges, but also automated search-based methods will need to be 10^6 faster to achieve what humans are already capable of involving visual 2D and 3D imagery pattern recognition.

Search-based methods have their origin in the field of database technology where query languages typically involve prepositional and/or predicate logic engines for deductive reasoning capability operating on data with either exact or greater-than/less-than condition. More recently, various methods have been added to such query languages to recognize patterns generated using a variety of statistical and other inductive reasoning capabilities. The latter problem, to discover a pattern from instances which are similar, is constrained by fundamental principles of randomness, noise, and independence. The search-based methods addressed herein complement statistical methods where interdependence between variables, linear and nonlinear, is assumed to be limited to a subrange of a variable(s), and therein, the task is to quickly identify and distinguish the subrange linear, nonlinear, and nonexistent relationships.

The research objective is to devise a computationally efficient pattern recognition capability to enable classification, functional mapping, and prediction. Of the various methods, neural nets [8], genetic algorithms [9], fuzzy logic, rough sets [10], and pyramidal nets-various combinations appear to show the greatest promise.

To be presented in this paper is the application of rough sets to materials data which builds upon an earlier publication [10]. The discussion that follows is intended to address the issue of rough sets as a method, the basic principles on which it has been constructed, and some of the open research problems associated with its development and application to materials discovery.

3. Rough sets

3.1. Origins of rough sets

Rough set theory [11] is a new mathematical approach to data analysis and data mining. After 15 years of pursuing rough set theory and its application the theory has reached a certain degree of maturity. In recent years we witnessed a rapid grow of interest in rough set theory and its application, world wide. Many international workshops, conferences and seminars included rough sets in their programs. A large number of high quality papers have been published recently on various aspects of rough sets.

The distinction between rough set theory and many other theories has been clarified. Particularly interesting is the relationship with fuzzy set theory and the Dempster-Shafer theory of evidence. The concepts of rough sets and fuzzy sets are similar but differ in their contribution to modeling with imprecision [12], i.e., a rough set establishes and qualifies the consistency of the relationship between the membership across two or more fuzzy sets. Therein, the relationship of rough sets with the theory of evidence is obviously more substantial [13]. In addition, rough set theory is related to discriminant analysis [14], Boolean reasoning methods [15] and others. The relationship between rough set theory and decision analysis is presented in [16,17]. Several extensions of the 'basic' model of rough sets have been proposed and investigated.

Various real life-applications of rough set theory have shown its usefulness in many domains. Very promising new areas of application of the rough set concept will emerge in the near future. These include rough control, rough data bases, rough information retrieval, rough neural networks and others, and it is clear, that rough set theory can contribute significantly to materials research.

3.2. Basic concepts

Rough set philosophy is founded on the assumption that with every object of the universe of discourse we associate some information (data, knowledge). Objects characterized by the same information are indiscernible (similar) in view of the available information about them. The indiscernibility relation generated in this way is the mathematical basis of rough set theory. Any set of all indiscernible (similar) objects is called an elementary set, and forms a basic granule (atom) of knowledge about the universe. Any union of some elementary sets is referred to as a crisp (precise) set-otherwise the set is rough (imprecise, vague).

Each rough set has boundary-line cases, i.e., objects which cannot be classified with certainty, by employing the available knowledge, as members of the set or its complement. Rough sets, in contrast to precise sets, cannot be characterized in terms of information about their elements. In the proposed approach, with any rough set is associated a pair of precise sets called the lower and the upper approximation of the rough set. The lower approximation consists of all objects which surely belong to the set, and the upper approximation contains all objects which possibly belong to the set. The difference between the upper and the lower approximation constitutes the boundary region of the rough set. These approximations are the two basic operations used in rough set theory.

Data are often presented as a table, columns of which are labeled by attributes, rows by objects of interest and entries of the table are attribute values. Such tables are

known as information systems, attribute-value tables, data tables, or information tables.

Usually we distinguish in information tables two kinds of attributes, called condition and decision attributes. Such tables are known as decision tables. Rows of a decision table are referred to as 'if...then...' decision rules, which give conditions necessary to make decisions specified by the decision attributes. An example of a decision table is shown in Table 1.

The table contains data concerning six cast iron pipes exposed to high pressure endurance tests. In the table *C*, *S* and *P* are condition attributes, displaying the percentage content of coal, sulfur and phosphorus respectively in pig-iron, whereas the attribute Cracks reveals the result of the test. The values of condition attributes are as follows (*C*, high)>3.6%, 3.5%≤(*C*, avg.)≤3.6%, (*C*, low)<3.5%, (*S*, high)≥0.1%, (*S*, low)<0.1%, (*P*, high)≥0.3%, (*P*, low)<0.3%.

The physical problem we are interested in is how the endurance of the pipes depends on the amount of *C*, *S* and *P* present in the pig-iron. In rough set terms, the problem is determining if there is a functional dependency between the decision attribute Cracks and the condition attributes *C*, *S* and *P*. In the language of rough set theory this leads to the question—given the set {2,4,5} of all pipes having no cracks after the test (or the set {1,3,6} of pipes having cracks), can cracks be uniquely defined (and thus predicted) in terms of these condition attribute values.

It can be easily seen that this is impossible, since pipes 2 and 3 display the same features in terms of attributes *C*, *S* and *P*, but they have different values of the attribute Cracks. Thus, information given in Table 1 is not sufficient to solve our problem. However, we can give a partial solution. We observe that if the attribute *C* has the value high for a certain pipe, then the pipe has cracks, whereas if the value of the attribute *C* is low, then the pipe has no cracks. Hence, employing attributes *C*, *S*, and *P*, we can say that pipes 1 and 6 surely are good, i.e., surely belong to the set {1,3,6}, whereas pipes 1, 2, 3 and 6 possibly are good, i.e., possibly belong to the set {1,3,6}. Thus, the sets {1,6}, {1,2,3,6}, and {2,3} are the respective lower, upper and boundary approximation region of the set {1,3,6}.

This means that the quality of pipes cannot be determined exactly by the content of coal, sulfur and phosphorus in the pig-iron, but can be determined only with some approximation.

Table 1
Example of an information system in the form of a decision table

Pipe	*C*	*S*	*P*	Cracks
1	High	High	Low	Yes
2	Avg.	High	Low	No
3	Avg.	High	Low	Yes
4	Low	Low	Low	No
5	Avg.	Low	High	No
6	High	Low	High	Yes

In fact, approximations determine the dependency (total or partial) between condition and decision attributes, i.e., they express a functional relationship between values of condition and decision attributes.

The degree of dependency between condition and decision attributes can be defined as a consistency factor of the decision table, which is the number of conflicting decision rules to all decision rules in the table. By conflicting decision rules, we mean rules having the same conditions but different decisions. For example, the consistency factor for Table 1 is $4/6=2/3$; hence the degree of dependency between cracks and the composition of the pig-iron is $2/3$. That means that four out of six (ca. 60%) pipes can be properly classified as good or not good on the basis of their composition.

We might be also interested in reducing some of the condition attributes, i.e., to know whether all conditions are necessary to make decisions specified in a table. To this end we will employ the notion of a reduct (of condition attributes). By a reduct we determine a minimal subset of condition attributes which preserves the consistency factor of the table. It is easy to compute, that in Table 1, we have two reducts: {*C*,*S*} and {*C*,*P*}. The intersection of reducts is called the core. In our example the core is the attribute *C*. This means, that in view of the data, coal is the most important factor causing cracks and cannot be eliminated from our considerations, whereas sulfur and phosphorus play a minor role and can be mutually exchanged as factors causing cracks.

Now we present the basic concepts more formally. Suppose we are given two finite, non-empty sets *U* and A, where *U* is the universe, and *A* a set of attributes. With every attribute $a \in A$ we associate a set V_a of its values, called the domain of *a*. Any subset *B* of *A* determines a binary relation $I(B)$ on *U* which will be called an indiscernibility relation, and is defined as follows:

$$x(B)y \text{ if and only if } a(x) = a(y) \text{ for every } a \in A,$$

where $a(x)$ denotes the value of attribute *a* for element *x*.

Obviously $I(B)$ is an equivalence relation. The family of all equivalence classes of $I(B)$, i.e., partition determined by *B*, will be denoted by $UB(B)$, or simply *UB*; an equivalence class of $I(B)$, i.e., block of the partition *UB*, containing *x*, will be denoted by $B(x)$.

If (x,y) belong to $I(B)$ we will say that *x* and *y* are *B*-indiscernible. Equivalence classes of the relation $I(B)$ (or blocks of the partition *UB*) are referred to as *B*-elementary sets. In the rough set approach the elementary sets are the basic building blocks of our knowledge about reality.

The indiscernibility relation will be used next to define basic concepts of rough set theory. Let us define now the following two operations on sets

$$B_*(X) = \{x \in U: B(x) \subseteq X\},$$

$$B^*(X) = \{x \in U: B(x) \cap X \neq \emptyset\},$$

assigning to every subset X of the universe U two sets $B_*(X)$ and $B^*(X)$ called the B-lower and the B-upper approximation of X, respectively. The set

$$BN_B(X) = B^*(X) - B_*(X)$$

will be referred to as the B-boundary region of X. If the boundary region of X is the empty set, i.e., $BN_B(X) = \varnothing$, then the set X is crisp (exact) with respect to B; in the opposite case, i.e., if $BN_B(X) \neq \varnothing$, the set X is referred to as rough (inexact) with respect to B.

A rough set can be also characterized numerically by the following coefficient

$$\alpha_B(X) = \frac{|B_*(X)|}{|B^*(X)|}$$

called accuracy of approximation, where $|X|$ denotes the cardinality of X. Obviously $0 \leq \alpha_B(X) \leq 1$. If $\alpha_B(X) = 1$, X is crisp with respect to B (X is precise with respect to B), and otherwise, if $\alpha_B(X) < 1$, X is rough with respect to B.

Approximation can be employed to define dependencies (total or partial) between attributes, reduction of attributes, decision rule generation and others. For details we refer the reader to the references. [15–33]

3.3. Applications

Rough set theory has found many interesting applications. The rough set approach seems to be of fundamental importance to AI and cognitive sciences, especially in the areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases, expert systems, inductive reasoning and pattern recognition. It seems of particular importance to decision support systems and data mining.

The main advantage of rough set theory is that it does not require any preliminary or additional information about the data to be analyzed, unlike probability in statistics, or basic probability assignment in Dempster-Shafer theory and grade of membership or the value of possibility in fuzzy set theory. The rough set theory has been successfully applied in many real-life problems in medicine, pharmacology, engineering, banking, financial and market analysis and others. Some exemplary applications are listed below.

There are many applications in medicine. In pharmacology the analysis of relationships between the chemical structure and the antimicrobial activity of drugs has been successfully investigated. Banking applications include evaluation of a bankruptcy risk and market research. Very interesting results have been also obtained in speaker independent speech recognition and acoustics. The rough set approach seems also important for various engineering applications, like diagnosis of machines using vibroacoustics symptoms (noise, vibrations) and process control. Application in linguistics, environment and databases are other important domains. First application of rough sets to material sciences, particularly interesting to this community, can be found in [10,18]. Rough set applications to materials research provides a new algorithmic method for autonomous discovery of relations between material properties and/or processing conditions, which can be very useful in designing new materials and/or their associated processing conditions [10]. More about applications of the rough set theory can be found in [19–25].

Application of rough sets requires suitable software. Many software systems for workstations and personal computers based on rough set theory have been developed; these include LERS [26], Rough DAS, Rough Class [27], and DATALOGIC [28]. Some of them are available commercially.

3.4. Example

A data set from Jackson et al. [32] was analyzed using rough sets. The purpose of this experiment was to determine if rough sets could predict behavior qualitatively similar to the curve fit obtained and demonstrated in [32]. Analysis was accomplished using a unique implementation of rough sets prepared by D. Ress [33].

3.4.1. Background information on the data set

The data set, as depicted in Table 2 and Fig. 2, consists of empirical data collected on a number of semiconductor and ionic compounds of interest because of their optical properties. The band gap range from 0.1 to about 10 eV represents wavelength transparencies ranging from the ultra violet (UV) to the far infrared (IR). Within this range there are a number of subranges related to atmospheric effects, absorption of solids, excitation of second harmonics, and other phenomena. The issue is to be able to choose a material based on knowledge of its band gap value that has good nonlinear optical properties. In [32] it is shown that there is a relationship between band gap and the nonlinear second-order optical coefficient ($\chi(2)$) of a material compound that is generally logarithmic, but curiously has two distinct behaviors. At low band gap values (<1.2 eV) the slope of the $\chi(2)$ vs gap curve is lower than for band gap values above 1.2 eV. Such behavior is not predicted theoretically from present models, and hence this empirical-derived relationship was something of a surprise. The curve is used for selecting candidate compounds based on their gap values.

Qualitatively, therefore, the behavior exhibited by the experimental data divides the compounds into two groups, one associated with each slope and the band gap dividing value of 1.2 eV. As a minimum, one would expect rough sets or any other classification scheme to obtain the same qualitative behavior.

The problem one has with such methods is determining the type of transformation of the numeric data into meaningful symbols, since rough sets operates on symbols with logical operators. The simplest approach is to take an

Table 2
Optical and band energy gap data [32] used in the rough sets analysis

Compound	Gap (eV)	$\chi(2)$ (pm V^{-1})^2
InSb	0.23	3234
Te	0.33	2581
GeSn	0.36	2308
InAs	0.36	838
$CdGeAs_2$	0.57	472
GaSb	0.72	1030
SiSn	0.84	1010
SiGe	0.9	674
SnC	1.2	556
$AgInSe_2$	1.2	7.6
InP	1.35	287
GaAs	1.4	180
CdTe	1.5	336
$CuInS_2$	1.53	14
$CuGaSe_2$	1.7	60
Se	1.7	258
$CdGeP_2$	1.72	218
$ZnSiAs_2$	1.74	146
$AgGaSe_2$	1.8	66
CdSe	1.8	108
Ag_3SbS_3	1.93	26
Ag_3AsS_3	2	50
GaSe	2.021	176
$ZnGeP_2$	2.05	150
GeC	2.1	76
HgS	2.1	110
$AgAsS_2$	2.14	50
b-SiC	2.26	60
GaP	2.3	210
ZnTe	2.3	184
$CuGaS_2$	2.43	20
CdS	2.485	88
$AgGaS_2$	2.638	36
ZnSe	2.7	156
AgI	2.8	16
CuBr	2.91	16
CuI	2.95	16
$CdGa_2S_4$	3.05	50
CuCl	3.17	14
ZnO	3.3	3.6
ZnS	3.9	74
$LiNbO_3$	4	10.9
KDP	7	1
SiO_2	8.4	0.8
InSe	1.25	248
GaS	2.5	270
map	2.6	12.6
pom	3	6
$LiIO_3$	4	10
urea	5.9	2
SiC	6	17.2
AlN	6.2	15
BBO	6.3	1.2
$HgGa_2S_4$	2.79	60

arbitrary number of intervals and divide the range of values into these intervals and assign a symbol to the interval (or 'bin'). Although this is acceptable, the transformation is not necessarily the best relative to an analysis. The bin ranges may be so gross as to mask any important behavior, or they may be so fine as to be useless. The analogical difficulty with neural nets training and generalization may be relevant here. If the number of bins is so large that every object is unique, then the dependence may be 100%, but the behavior is trivial. Reducing the number of bins will change the dependence, because the numeric data is 'blurred' as the number of bins decreases. At the other extreme of 1 bin, the dependence is also 100%, and is also a trivial result. So the task is to find a set of bins between these extremes.

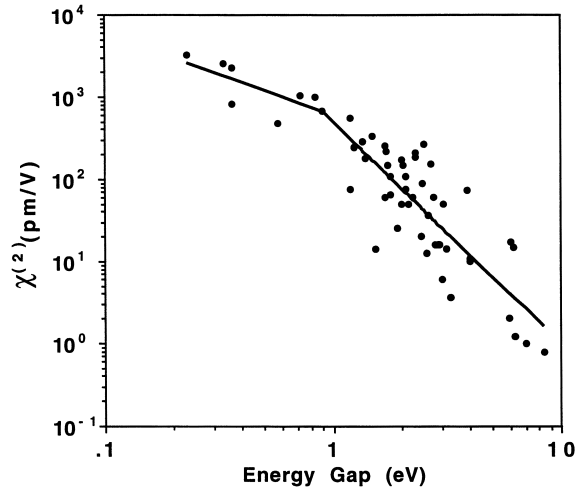

Fig. 2. Plot of $\chi(2)$ against energy gap values for compounds listed in Table 2 (From ref. [1]].

This problem has been approached by use of genetic algorithms to search the possible bin sets to find those values which will produce the highest dependence. Such automatic search of the bin space is efficient in the sense that the user is not required to choose a bin set based on intuition or previous knowledge.

Using a combined rough set-evolutionary method, a genetic algorithm used in conjunction with rough sets [33] identified relatively high dependencies for a two-bin dependent variable and seven-bin independent variable. The dependence was relatively low (0.33), which is somewhat troublesome for interpretation purposes. However, the selection of 2 bins for $\chi(2)$ is appropriate, since this reflects the two-slope behavior seen in the experimental data. The number of bins selected for band gap was 7, a value not unreasonable. It is believed by the authors, that the lack of a high dependence value is a consequence of the scatter in the data set.

To illustrate how scatter or 'noise' affects the dependence, a simple two-slope function was created which mimics the behavior of the experimental data (Fig. 3). With no noise present, the dependence was 1. This is an expected result and both validates the rough set program and lends evidence to our hypothesis. Adding random noise to the function induces a spread in the values of the function and reduces the dependence by spreading the standard deviation. The accuracy for these four runs degraded from absolute certainty for the no or low noise

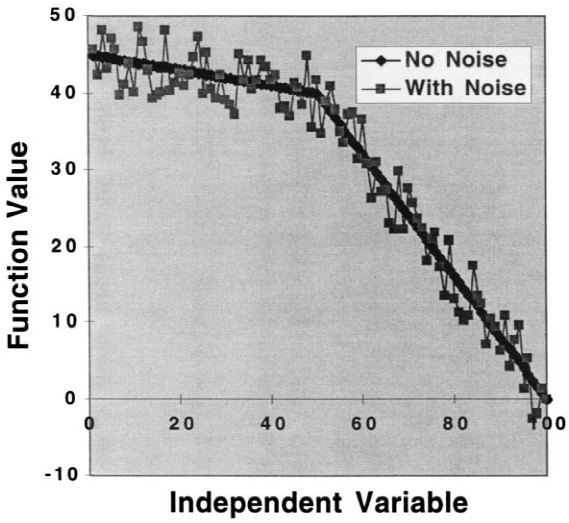

Fig. 3. Plot of test function with exact (solid line) and noisy values. The function is designed to have the same behavior as the experimental data analyzed by rough sets.

case to 0.20 accuracy for the large noise case. In each case the number of bins selected as best for the dependent variable was two. The number of bins selected for the independent variable depended on the noise level. For low noise the bin number was 18, dropping to 14 for the next higher noise, and rising again to 18–19 for the high to very high noise levels. Thus, the RS method successfully identifies the presence of two slopes in the curve (2 bins), the essential feature of this function. The number of bins chosen for the dependent variable (14–19) samples the range satisfactorily.

The behavior of this simulation provides some insight into the electro-optical data results, suggesting that the low accuracy is connected to the spread in the values of the data. But the essential feature of the $\chi(2)$ data, the break in the slope at about 1.2 eV, is reproduced in the RS results. Hence, RS is useful for identifying such aspects of behavior in this data.

4. Discussion

The rough sets approach to data analysis has many important advantages:

- Provides an efficient, and when combined with a genetic algorithm, an autonomous method for finding patterns in data.
- Identifies relationships, although qualitative that include nonlinear, that would not be found using statistical methods.
- Allows both qualitative and quantitative data to be analyzed.
- Finds minimal sets of (independent data to predict dependent) data, i.e., data reduction.
- Evaluates the significance (distinguishing between levels of consistency) of data.
- Generates sets of decision rules (based on relations) from data.
- Is easy to understand (does not require mathematics background), e.g., statistics.
- Offers straightforward (lower vs. upper approximation) interpretation of results.
- Well suited for parallel processing.

Although rough set theory has many achievements to its credit, there are, nevertheless, several theoretical and practical problems which require further attention. Particularly noteworthy is the lack of widely accessible software for rough set based data analysis, particularly for large collections of data.

Despite the many valuable methods of decision rule generation (ID3, pyramidal networks, clustering algorithms, fuzzy sets, Dempster-Shafer, etc.), to include rough set theory, more research is needed, particularly when quantitative attributes are involved. In this context, new discretization methods for quantitative attribute values are badly needed. Also, an extensive study of a new approach to missing data is very important. Comparison to other similar methods still requires due attention, although important results have been obtained in this area. Particularly interesting is a study of the relationship between neural networks and the rough set approach to feature extraction from data. Last but not least, a rough set computer is badly needed for more serious applications. Some research in this area is already in progress. For further investigation of the fundamental theory and basic concepts of rough set theory, the reader is referred to [17,29–31].

5. Conclusions

The rough sets approach to data analysis provides a method for analysis of materials data that has some distinct advantages, particularly when used in conjunction with other data analysis approaches. Search for hidden patterns in data can be accomplished efficiently in terms of the algorithms of rough sets. Relationships present in the data can be revealed using qualitative representations as well as numerical. Reduction of the number of variables of importance for a particular class of objects is possible, an important advantage when experimental materials systems involve a large number of variables. A direct result of rough sets analysis is an evaluation of the significance of the data, and also the generation of rules suited to decision making about the data. As a preprocessing tool, rough sets, therefore, offers good potential for materials design and process design.

Acknowledgements

AGJ acknowledges the support provided under AF Contract F33615-94-D-5801 with the Wright Laboratory, Materials Directorate. The authors express their thanks to Dr. W. Ziarko for discussions on rough sets. Also, our thanks to D. Ress for use of the rough sets analysis application developed by him. Z. Pawlak gratefully acknowledges the support of the Air Force Contract F6 1 708-97-WO 196.

References

[1] H. Koinuma, Why Crystal Engineering of Oxides?, MRS Bulletin, Sept. 1994, pp 21–24.
[2] D. Dorsey, Internal Memoranda on New World Vistas, May 1995.
[3] A.I. Adamatzky, Mathematical Computer Modeling 23(4) (1996) 51–56.
[4] Visualization Tools Linked to Data Acquisition and Ease of Use, R and D Magazine, December, (1995) 43–44.
[5] M.F. Ashby, Materials Selection in Mechanical Design, Materials and Process Selection Charts, Pergamon Press, London, 1992.
[6] N. Kiselyova, J. Alloys Comp. 197 (1993) 159–165.
[7] E. Savitskii, V.B. Gribulya, N.N. Kiselyova, M. Ristich. Z. Nikolich, Z. Stoyilkovich, M. Zhivkovich, I.P. Arsenteva, [Prediction of Material Properties using IBM Computers], Pro gnoztrovanije b Materialovedeniij c primeneniem EBM, Nauka Moskva, 1990.
[8] S. Thaler, Neural Networks that Autonomously Create and Discover, unpublished paper, 1995.
[9] D. Wood and J. Park, Discovery Systems for Manufacturing, Wright Laboratory Technical Report, WL-TR-94-4008, January 1994.
[10] A.G. Jackson, S.R. LeClair, M.C. Ohmer, W. Ziarko, H. Al-Kamliawi, Rough Sets Applied to Materials Data, Acta Metallurgica et Materialia 44(11) (1996) 4475–4484.
[11] Z. Pawlak, Int. J. Computer Info. Sci. 11 (1982) 341.
[12] Z. Pawlak, A. Skowron, in: R.R. Yaeger, M. Fedrizzi, J. Kacprzyk (Eds.), Advances in the Dempster Shafer Theory of Evidence, John Wiley and Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore, 1994.
[13] A. Skowron, J.W. Grzymala-Busse, in: R.R.M. Fedrizzi, J. Kacprzyk (Eds.), Advances in the Dempster-Shafer Theory of Evidence, John Wiley and Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore, 1994.
[14] E. Krusifiska, R. Slowinski, J. Stefanowski, Applied Stochastic Models Data Analysis 8 (1992) 43.
[15] A. Skowron, C. Rauszer, in: R. Slowinski (Ed.), Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, Kluwer Academic Publishers, Boston, London, Dordrecht, 1992, p. 471.
[16] Z. Pawlak, R. Slowinski, Eur. J. Oper. Res. 72 (1994) 443.
[17] R. Slowinski, AI Expert 10 (1995) 18.
[18] A.G. Jackson, M. Ohmer, H. Al-Kamhawi, in: T.Y. Lin (Ed.), The Third International Workshop on Rough Sets and Soft Computing Proceedings (RSSC'94), San Jose State University, San Jose, California, USA, 1994.
[19] T.Y. Lin, N. Cercone, Rough Sets and Data Mining - Analysis of Imperfect Data, Kluwer Academic Publishers, Boston, London, Dordrecht, 1997, p. 430.
[20] T.Y. Lin, A.M. Wildberger, The Third International Workshop on Rough Sets and Soft Computing Proceedings RSSC'94), San Jose State University, San Jose, California, USA, 1995.
[21] R. Slowinski, Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, Kluwer Academic Publishers, Boston, London, Dordrecht, 1992, p. 471.
[22] 5. Tsumoto, S. Kobayashi, T. Yokomori, H. Tanaka, A. Nakamura, Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets and Machine Discovery, The University of Tokyo, 1996, p. 465.
[23] P.P. Wang, Proceedings of the Second Annual Joint Conference on Information Sciences, Wrightsville Beach, North Carolina, USA, 1995.
[24] P. Wang, Joint Conference of Information Sciences, Vol. 3, Rough Sets and Computer Sciences, Duke University, 1997, p. 449.
[25] W. Ziarko, Rough Sets, Fuzzy Sets and Knowledge Discovery. Proceedings of the International Workshop on Rough Sets and Knowledge Discovery (RSKD'93), Banff, Alberta, Canada, October 12–15, Springer-Verlag, Berlin, 1993, p. 476.
[26] J.W. Grzymala-Busse, in: R. Slowinski (Ed.), Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, Kluwer Academic Publishers, Boston, London, Dordrecht, 1992, p. 471.
[27] R. Slowinski, J. Stefanowski, in: R. Slowinski (Ed.), Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, Kluwer Academic Publishers, Boston, London, Dordrecht, 1992, p. 471.
[28] A. Szladow, PC AI 7(1) (1993) 40.
[29] Z. Pawlak, Rough Sets-Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Boston, London, Dordrecht, 1991, p. 229.
[30] Z. Pawlak, J.W. Grzymala-Busse, R. Slowinski, W. Ziarko, Commun. ACM 38 (1995) 88.
[31] A. Szladow, W. Ziarko, AI Expert 7 (1993) 36.
[32] A.G. Jackson, M.C. Ohmer, S.R. LeClair, IR Phys. Techn. 38 (1997) 233–244.
[33] D. Ress, North Carolina State University, 1997, Application in Hypercard for analysis of rough sets and selection of optimum bins to produce highest dependency.